УДК 517.958

А. В. Абрагин

ПРИМЕНЕНИЕ КОНФЛЮЕНТНОГО АНАЛИЗА В МЕТОДЕ ГРУППОВОГО УЧЕТА АРГУМЕНТОВ

Приведен классический многорядный алгоритм группового учета аргументов, рассмотрено влияние погрешностей на параметры полученной математической модели. Предложена модификация алгоритма, позволяющая учесть погрешности данных и избежать возрастания неопределенности параметров модели на каждом ряде. Предложен также новый критерий качества модели, учитывающий неопределенность параметров.

Цель настоящей работы — определить область применимости метода группового учета аргументов (МГУА) при наличии погрешности исходных данных как на входе, так и на выходе исследуемой системы. Будем полагать плотности распределения вероятностей данных известными. В настоящей работе ограничимся изучением случая нормального распределения.

Рассмотрим алгоритм МГУА, изложенный в работах [1, 2], с помощью которого решается следующая задача.

Требуется найти аналитическую модель многопараметрического процесса, которая будет использована для прогнозирования или идентификации данного процесса.

Пусть в общем виде модель представлена следующим образом:

$$y = h(x), \quad x \in V,$$

где y — вектор наблюдений исследуемого процесса; x — вектор параметров, от которых зависит прогнозируемая величина; V — область допустимых значений параметров.

Параметрическое семейство отображений F, зависящих от параметра Θ, назовем структурой отображения [1]. Структуру F назовем несмещенной оценкой отображения h, если

$$F(x; M[\hat{\Theta}]) = h(x), \quad x \in V,$$

где $\hat{\Theta}$ — любая несмещенная оценка параметра Θ; $M[\cdot]$ — математическое ожидание [1].

Пусть выполнены следующие предположения:

1) существует единственная зависимость h между входной $x = (x_1, x_2, \ldots, x_m)$ и выходной $y = h(x), x \in V$, величинами (величинами на входе и на выходе исследуемой системы);

2) задан класс Ф — класс структур отображений;

3) задана детерминированная матрица X размерности $(n \times m)$ значений входной величины в n точках;

4) задан n-мерный вектор Y наблюдений выходной величины, причем выходная величина наблюдается с ошибкой $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^{\mathrm{T}}$, величины $\varepsilon_i, i = 1, 2, \ldots, n$, полагаются случайными независимыми одинаково распределенными величинами с нулевыми математическими ожиданиями и конечной дисперсией.

В этих предположениях требуется найти структуру $F \in \Phi$, являющуюся несмещенной оценкой отображения h.

Для решения поставленной задачи необходимо:

1) сформировать конечное множество отображений $F \in \Phi$;

2) вычислить параметры каждого отображения и получить несмещенные оценки истинного отображения;

3) пользуясь критериями отбора, выбрать лучшую модель.

Рассмотрим пункт 1) решения. Множество F можно сформировать различным образом в зависимости от исходных предположений об исследуемом процессе; на практике чаще всего используются полиномы многих переменных. В простейшем случае множество F может быть множеством всех полиномов степени, не превосходящей некоторое заранее заданное число. Этот подход может быть реализован для задач с небольшим числом входных данных, но при количестве входных переменных более трех полнопереборный алгоритм требует значительных вычислительных затрат [2].

Для снижения вычислительных затрат используются так называемые многорядные алгоритмы МГУА, в которых множество F не формируется сразу, а создается на каждом этапе; при этом модели, признанные лучшими на первом этапе, являются основой для формирования множества моделей следующего этапа. Эти алгоритмы относятся к алгоритмам генетического поиска [3], они известны в теории нейросетей как “полиномиальные сети” [4].

Пусть исследуемый процесс описывается некоторой функцией $f(x_1, x_2, \ldots, x_m)$.

На первом этапе строим множество моделей вида

$$y_1 = f(x_1, x_2), \quad y_2 = f(x_2, x_3), \quad \ldots, \quad y_{m-1} = f(x_{m-1}, x_m).$$

Далее вычисляем их параметры и выбираем модели по некоторому критерию.

На втором этапе модели служат основой для формирования множества моделей второго этапа вида

$$z_1 = f(y_1, y_2), \quad z_2 = f(y_2, y_3), \quad \ldots, \quad z_{m-2} = f(y_{m-2}, y_{m-1}).$$

Процесс повторяется до тех пор, пока значение критерия отбора не перестанет уменьшаться. Этот метод, более подробно изложенный в работах [1, 2], позволяет сократить время вычислений.

Рассмотрим пункт 2) решения. Вычисление параметров каждой модели в работах [1, 2] предлагается проводить методом наименьших квадратов. Однако, как показано в работах [5, 6], с помощью метода наименьших квадратов не всегда можно получить несмещенную оценку параметров. Как показывают вычислительные эксперименты, для данных, имеющих погрешность на каждом этапе, происходит возрастание неопределенности параметров. Эксперимент проводился следующим образом: были выбраны полином пятой степени и 20 точек, принадлежащих графику некоторой функции. К координатам точки была добавлена нормально распределенная погрешность. Эта операция была проделана 100 раз с одинаковыми параметрами нормального распределения. Далее были проведены расчеты множества моделей-претендентов — полиномов степени, не превышающей два. Отбор лучшей модели проводился по критерию регулярности. Затем были проведены расчеты для второго и последующих этапов (количество рядов не ограничивалось). В результате истинная модель была найдена 42 раза, при этом значения целевой функции критерия регулярности принципиально не различались, что говорит о существенном влиянии погрешностей входных и выходных данных на выбор модели. В остальных случаях лучшими моделями являлись полиномы достаточно высокой степени (до 25), которые при таком количестве данных могут точно аппроксимировать кривую, проходящую через точки, и прогностическая ценность которых невелика.

В настоящей работе предлагается для получения несмещенной оценки использовать метод конфлюентного анализа, изложенный в работах [5, 6]. Рассмотрим более подробно алгоритм получения оценок.

Пусть отображение f имеет следующий вид:

$$f(x_1, x_2, \ldots, x_m) = \sum_{l=1}^{k} \theta_l \prod_{j=1}^{m} x_j^{\alpha_{jl}}.$$

Тогда в соответствии с работой [5] несмещенную оценку параметра Θ можно получить минимизацией функционала

$$L = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{(x_{ij}-\xi_{ij})^2}{\sigma^2(x_{ij})} + \frac{(y_i - f(\xi_{ij},\Theta))^2}{\sigma^2(y_i)}\right),$$

где $\sigma^2(x_{ij})$, $\sigma^2(y_i)$ — дисперсии соответствующих переменных; x_{ij} — значение i-й переменной в j-м наблюдении; ξ_{ij} — искомые значения переменных; $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$.

В работе [5] задача сводится к минимизации квадратичной формы

$$G(\Theta) = \frac{1}{2}\Theta A\Theta + a^{\mathrm{T}}\Theta,$$

где

$$A = [A_{rp}], \quad A_{rp} = \sum_{i=1}^{n}\frac{1}{\sigma^2(y_i)}\prod_{j=1}^{m}x_{ij}^{\alpha_{jl}}, \quad r = 1,2,\ldots,k, \quad p = 1,2,\ldots,k,$$

$$a = (a_1, a_2, \ldots, a_r), \quad a_r = -\sum_{i=1}^{n}\frac{y_i}{\sigma^2(y_i)}\prod_{j=1}^{m}x_{ij}^{\alpha_{jl}}, \quad r = 1,2,\ldots,k.$$

Точные значения входных переменных вычисляются при решении уравнения

$$\frac{x_{ij}-\xi_{ij}}{\sigma^2(x_{ij})} + \frac{y_i - f(\xi_{ij},\Theta)}{\sigma^2(y_i)}\frac{\partial f}{\partial \xi_{ij}} = 0, \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,m,$$

где

$$\frac{\partial f}{\partial \xi_{ij}} = \sum_{i=1}^{k}\theta_l\alpha_{ij}x_{i1}^{\alpha_{1l}}\ldots x_{ij}^{\alpha_{jl}-1}\ldots x_{im}^{\alpha_{ml}}.$$

Полученные решения должны удовлетворять условию

$$|x_{ij} - \xi_{ij}| \leqslant 3\sigma(x_{ij}).$$

Элементы дисперсионной матрицы находятся путем вычисления матрицы, обратной

$$M = [M_{rp}], \quad M_{rp} = -\frac{\partial^2 F}{\partial\theta_r\partial\theta_p} \quad \text{при} \quad \Theta = \hat{\Theta},$$
$$r = 1,2,\ldots,k, \quad p = 1,2,\ldots,k,$$

где $\hat{\Theta}$ — найденные оценки параметров Θ.

Применение конфлюентного анализа в МГУА обеспечивает учет погрешностей входных и выходных данных и позволяет получить дисперсионную матрицу.

Рассмотрим пункт 3) решения. Критерии отбора, используемые в алгоритмах МГУА, получены на основе теоремы Гёделя о неполноте и принципа внешнего дополнения. Принцип получения критериев следующий: экспериментальные данные разбиваются на две части — на множества A и B; множество A используется для вычисления параметров модели, а множество B (внешнее дополнение) используется для анализа полученной модели.

Наибольшее распространение получили такие внешние критерии, как критерий регулярности и критерий минимума смещения [2]. Первый можно рассматривать как простейшую оценку эффективности структуры модели. Рассмотрим эти критерии подробнее.

Пусть оценивается качество структуры f. Найдем оценку параметра Θ на основании информации, содержащейся в множестве A. Обозначим такую оценку параметра Θ_A. Рассмотрим вектор невязок модели $f(\cdot;\Theta_A)$ на множестве B. Он равен $Y_B - F(X_B;\Theta_A)$, где

$$F(X_B;\Theta_A) = \begin{bmatrix} f(x_{n_A+1,1}, & x_{n_A+1,2}, & \dots, & x_{n_A+1,m}; & \Theta_A) \\ f(x_{n_A+2,1}, & x_{n_A+2,2}, & \dots, & x_{n_A+2,m}; & \Theta_A) \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ f(x_{n,1}, & x_{n,2}, & \dots, & x_{n,m}; & \Theta_A) \end{bmatrix};$$

здесь X_B, Y_B — элементы внешнего дополнения; n_A — количество элементов в множестве A.

Целевой функцией критерия регулярности является евклидова норма указанного вектора невязок [1]:

$$\Delta^2(B) = (Y_B - F(X_B;\Theta_A))^{\mathrm{T}}(Y_B - F(X_B;\Theta_A)). \qquad (1)$$

Целевая функция критерия минимума смещения определяется следующей формулой [1]:

$$n^2_{\text{см}} = (F(X;\Theta_A) - F(X;\Theta_B))^{\mathrm{T}}(F(X;\Theta_A) - F(X;\Theta_B)), \qquad (2)$$

где Θ_B — оценка параметра Θ, полученная на множестве B.

Отметим случайность невязки. В настоящей работе в качестве целевой функции критерия предлагается вероятность того, что значения вектора Y не будут выходить за пределы толерантного интервала, вычисленного для каждой точки. Тогда аналог целевой функции критерия (1) будет иметь вид

$$C_1 = \prod_{i=1}^{n} \mathrm{P}\{|Y_{Bi} - F(X_{Bi};\Theta_A)| < \eta\}, \qquad (3)$$

где η — допустимая ошибка.

Аналогичные рассуждения применяются при модификации целевой функции критерия (2):

$$C_2 = \prod_{i=1}^{n} \mathrm{P}\{|F(X_i;\Theta_A) - F(X_i;\Theta_B)| < \eta\}. \qquad (4)$$

Вероятности в формулах (3) и (4) рассчитываются исходя из методики, изложенной в работе [5]: если число точек достаточно велико и можно считать, что оценки параметров распределены относительно их математических ожиданий по нормальному закону, то для получения толерантных интервалов применяют безразмерную t-статистику Стьюдента, которая подчиняется t-распределению с $n-2$ степенями свободы. Таким образом, практическое вычисление целевой функции критерия (3) затруднено, а целевая функция критерия (4) вычисляется как вероятность одновременного наступления n событий — принадлежности параметров модели заданному интервалу. В дальнейших численных экспериментах будем использовать целевую функцию критерия (4).

Таким образом, можно реализовать следующий модифицированный алгоритм МГУА.

1. Формируем первый ряд моделей-претендентов и вычисляем их коэффициенты по методу из работы [5].

2. Вычисляем значения целевых функций критериев (3) или (4) и сортируем модели.

3. На основании выбранных лучших моделей строим новое множество моделей-претендентов, осуществляем замену переменных и вычисляем коэффициенты моделей-претендентов.

4. Вычисляем значения целевых функций критериев (3) или (4) и сортируем модели.

5. Повторяем пункты 3 и 4 до тех пор, пока значения целевых функций критериев не начнут уменьшаться или вычисления не будут прекращены из-за ограничений параметров алгоритма (допустимого количества слоев, сложности модели, ограниченности времени вычислений).

Для проверки положений, изложенных в настоящей работе, проведены численные эксперименты. Выбрана функция с существенной нелинейностью, притом достаточно простая: $f(x) = x^5 - x^3 + x - 1$. Получено 20 точек, принадлежащих графику данной функции. К координатам точек добавлена нормально распределенная погрешность; это проделано 100 раз с одинаковыми параметрами нормального распределения. Далее проведен расчет множества моделей, вычислены значения целевой функции критерия (4), предлагаемой в настоящей работе, и на основании критерия выбраны лучшие модели. Проведено 100 таких

расчетов — для каждого варианта экспериментальных данных. Анализ показал, что при использовании предлагаемых критериев верные результаты были получены в 69 случаях — чаще, чем при использовании критерия регулярности (в 42 случаях), при этом количество полиномов высокой степени было меньше.

Эксперименты показали, что применение критериев с модифицированными целевыми функциями позволяет выбрать модели с учетом неопределенности их параметров (неопределенность учитывается при вычислении вероятностей в формулах (3), (4)) и избежать появления чрезмерно усложненных моделей.

Преимуществом предложенных критериев является возможность учета неопределенности как входных, так и выходных данных, соответствие этих критериев наиболее распространенной формулировке задачи моделирования сложных процессов (“требуется получить модель, которая описывает экспериментальные данные с заданной точностью и заданным уровнем значимости”), а также отсутствие необходимости вводить усложненные критерии отбора, использующие “штраф за сложность модели”, как, например, в работе [8].

СПИСОК ЛИТЕРАТУРЫ

1. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. – М.: Радио и связь, 1987.
2. Ивахненко А. Г. Долгосрочное прогнозирование и управление сложными системами. – Киев: Техника, 1975.
3. Ivakhnenko A. G. Recent Developments of Self-Organising Modeling in Prediction and Analysis of Stock Market // Pattern Recognition and Image Analysis. – 1995. – V. 5. – № 4. – P. 527–535.
4. Ivakhnenko A. G., Ivakhnenko G. A., Muller J. A. Self-Organisation of Neuronets with Active Neurons// Pattern Recognition and Image Analysis. – 1994. – V. 4. – № 2. – P. 177–188.
5. Грешилов А. А. Математические методы построения прогнозов. – М.: Радио и связь, 1997.
6. Грешилов А. А. Статистические методы принятия решений с элементами конфлюэнтного анализа. – М.: Радио и связь, 1998.
7. Вентцель Е. С. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1998.
8. Техническая документация к программе Neuro Shell. – Ward System Group, Inc.

Статья поступила в редакцию 1.04.2003

Артур Викторович Абрагин родился в 1973, окончил в 1996 г. МГТУ им. Н.Э. Баумана. Ассистент кафедры “Вычислительная математика и математическая физика” МГТУ им. Н.Э. Баумана. Автор 5 научных работ в области нейросетей, распознавания образов.

A.V. Abragin (b. 1973) graduated from the Bauman Moscow State Technical University in 1996. Assistant of "Computation Mathematics and Mathematical Physics" department of the Bauman Moscow State Technical University. Author of 5 publications in the field of neuronets, pattern recognition.